基于大语言模型少样本学习的领域知识图谱构建方法研究

——以图书馆学领域为例

【摘要】：文章探讨了利用大预言模型少样本学习的方式构建领域知识图谱，并以图书馆学领域为例进行了实验，旨在减少人工干预,提高构建效率,并探索其在不同领域的普适性。实验选取了四种商用语言模型(GPT-4.0、Claude-3-Opus、Claude-3.5-sonnet 和 Gemini-1.5-pro),通过多次提示词设计和知识抽取,构建了图书馆学领域的核心概念和等级体系及其类属性，进而构建图书馆学领域知识图谱。结果表明大语言模型可以降低人工干预，提高知识图谱构建的效率和一致性，同时增强其可扩展性和泛化能力。然而，提示词设计对于模型效果至关要，可能需要领域知识的补充。

关键词:大语言模型,少样本学习,知识图谱,图书馆学,提示词设计

Research on Domain Knowledge Graph Construction Method Based on Large Language Modeling with Less Sample Learning

--Taking the field of librarianship as an example

[**Abstract]**The article explores the construction of domain knowledge graphs by using large prediction models with fewer samples learning, and conducts experiments in the field of librarianship as an example, aiming to reduce manual intervention, improve the construction efficiency, and explore its universality in different domains. Four commercial language models (GPT-4.0, Claude-3-Opus, Claude-3.5-sonnet and Gemini-1.5-pro) are selected for the experiments, and the core concepts and hierarchies of librarianship and their class attributes are constructed through multiple cue word design and knowledge extraction, and then the librarianship domain knowledge graph is constructed. The results show that the large language model can reduce manual intervention, improve the efficiency and consistency of knowledge graph construction, and enhance its scalability and generalization ability. However, cue word design is critical to the modeling effect and may need to be supplemented by domain knowledge.
[Keywords]Big Language Modeling;Less Sample Learning;Knowledge Graph;Librarianship;Cue word design

1 引言

随着数字化人工智能的迅猛发展，知识图谱作为一种强大的知识管理和智能应用技术，在国家层面的应用日益广泛，涵盖了科研、公共服务、智慧城市、农业、医疗等多个领域[1]。通过对知识进行结构化表示和图形化展示，知识图谱不仅能够帮助用户全面、有效地理解知识之间的层次和关系体系，还能辅助挖掘用户潜在的知识需求[2][3]。因此，构建领域知识图谱，将海量的领域知识数据进行高效抽取、融合、存储、推理，有助于知识的高效利用，从而降低用户因知识储

备不足带来的检索负荷[4]，也为未来实现知识共享、跨领域知识检索和智能问答打下良好的基础[5]。

然而，随着知识图谱规模的扩大，传统的图谱构建和补全技术面临着诸多挑战，例如数据获取、实体识别、知识抽取和实体消歧等问题[6-9]。传统的知识图谱构建方法往往依赖大量的人工参与，例如制定规则、构建模板、标注数据等，成本高昂且效率低下成本高昂且效率低下[10]。在处理大规模、异构知识数据上，传统方法的可扩展性比较差，无法适应知识快速增长的需求。

2023 年 4 月 OpenAI 发布了最新一代大语言模型 ChatGPT-4，标志着大语言模型技术的进一步升级，也预示着自然语言处理领域的重大进步[1]。ChatGPT-4 在语言理解与生成、多模态处理、少样本学习、模型优化、应用场景拓展以及伦理与安全等方面的显著突破，为广泛的实际应用和未来研究提供了强大的工具和新方向，也为人工智能在各个领域将带来更多的创新和变革。

大语言模型的兴起与生成式人工智能（AIGC）的应用，使得各领域研究热点从传统的“预训练-微调”范式转向“提示工程-Prompt”范式[11][12]。这一转变不仅能够更好地弥补传统知识图谱构建方法的不足，还为知识图谱构建方法开辟了新思路。越来越多的研究表明，将生成式大语言模型（Large Language Models, LLMs）与知识图谱相结合的“图模互补”形式应用到各个领域，可以很好地利用两者间的优势互补来增强知识图谱的补全性能[10][13]。文献[14-15]着重研究基于提示学习的关系抽取问题，文献[16-17]就基于文本生成的知识图谱补全问题分别提出了基于 BERT 模型和自动生成提示补全方法。文献[18]就基于元学习的知识图谱构建进行了研究，从而提高了模型在学习不同关系上的泛化能力。

基于上述背景，本文以图书馆学领域知识图谱的构建方法为例，探索如何通过设计有效的提示词(prompts)来引导大语言模型生成和整合特定领域知识。这种方法不仅可以快速适应不同领域的知识需求，还能通过交互式的方式不断改进、扩展和深化知识图谱。研究旨在利用大语言模型的少样本学习能力，减少人工干预，探索领域知识图谱的高效构建的一致性和普适性方法。

2 构建流程与研究方法

2.1 构建流程

领域知识图谱构建的核心流程主要包括以下步骤：分析目标领域，梳理关键知识，明确概念、层次结构及关系属性、概念的收集与融合、数据存储与图构建。具体构建流程参考图 1。

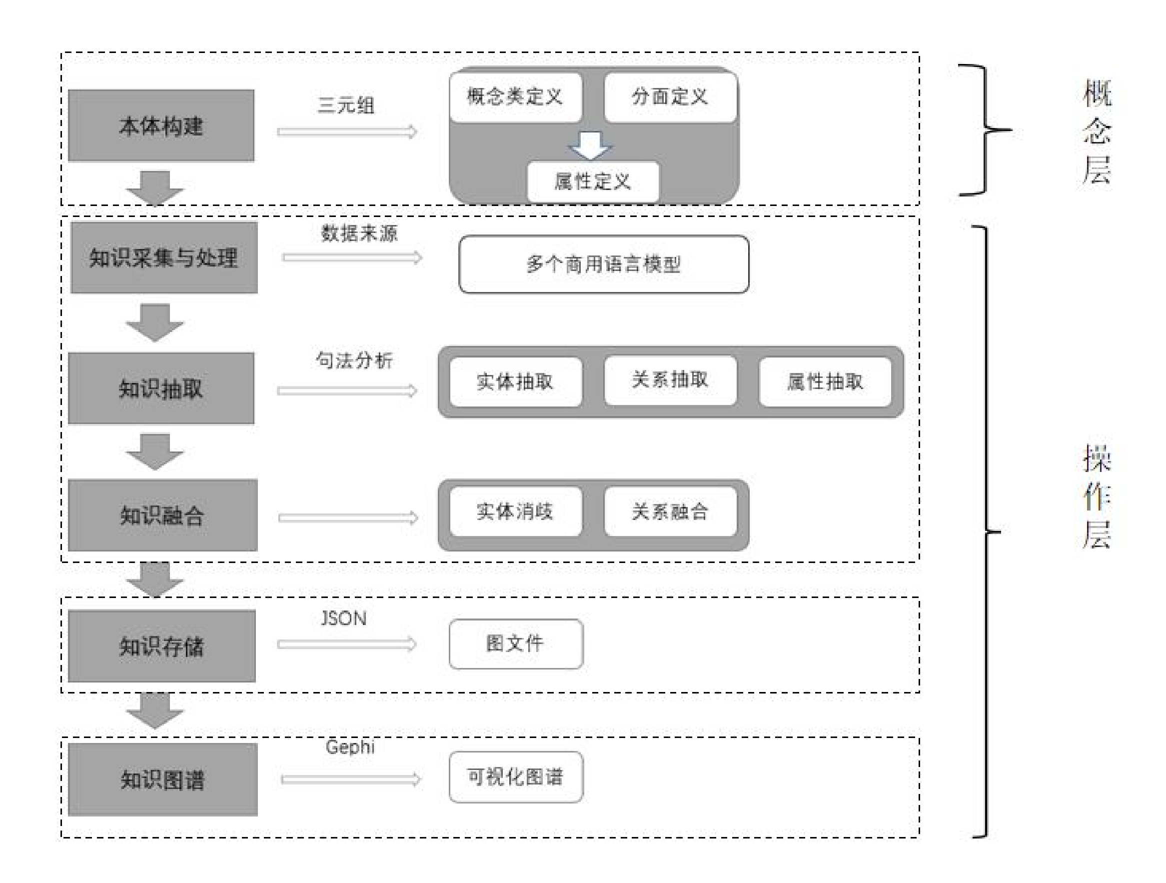


图 1 领域知识图谱构建流程总图

Fig. 1 General diagram of domain knowledge graph construction process

在整个知识图谱构建过程中，领域知识的本体构建是至关重要的一环，决定了领域知识图谱构建的准确与否。本体构建方法主要包括以下几个步骤：

(1)领域分析：确定领域本体的范围和目标，研究领域特征和专业术语。

(2)概念提取：以提示词文本从选择的商用模型中抽取、收集领域术语。提示文本可以通过多次调整文本格式，获取需要的领域知识概念。

（3）概念层次化：通过提示文本方式建立概念之间的上下位关系，构建概念分类体系。

（4）关系定义：确定概念间的语义关系，定义属性和关系的类型。

（5）融合处理：对于多次抽取出的概念，根据其权重进行处理，对于相同或相似进行整合、对于概念的差异进行消歧、对于概念的关系或属性进行统一等处理。

在完成知识抽取、融合任务后，将知识数据的实体、属性和关系以“节点—边”的形式借助于 JSON，以.GEXF 文件格式储存到图数据库，其中节点用于存储实体，而有向边则用于存储实体属性和关系，这种以图形的形式来存储复杂的知识，具有很强的扩展性和直观性。最后再借助于 Gephi 可视化分析软件生成所需要的领域知识图谱。

2.2 研究方法

少样本学习（Few-Shot Learning，FSL）是一种机器学习技术，旨在从少量训练样本中学习有效的模型。利用语言模型进行少样本学习在自然语言处理(NLP)领域已经取得了显著进展。特别是预训练的语言模型（如 GPT-4、BERT、T5 等）在少样本学习中表现出了强大的能力。预训练语言模型的少样本学习最重要的一

个策略就是通过构造适当的文本提示（prompts），使预训练语言模型能够在少量样本的情况下执行各种任务。领域知识图谱的构建主要核心在于如何构建高效的文本提示，通过提示词的文本设计，使得模型能够准确理解并回应用户的意图。

少样本学习主要涉及提示词设计、上下文设置以及示例引导三方面内容。

(1) 提示设计

构建一个能有效引导模型生成所需输出的输入提示。这包括选择合适的词汇、句子结构以及明确的问题或指示。这里必须确保提示文本清晰、具体并易于理解，避免模糊的描述和歧义。如果面对复杂的问题时，可以分步指示，将复杂问题分解为多个步骤，每个步骤提供明确的指示。

(2)上下文设置

为模型提供必要的上下文信息，使其能够更好地理解并回应请求。为了便于更好的被语言模型理解，在提供上下文时，需要在提示中包含相关背景信息，以帮助模型理解问题的背景和细节。

(3)示例引导

示例引导主要是通过提供示例来展示期望的输出格式和内容。

示例 1：

提示	作为图书馆学领域专家，请列出领域最核心的 10 个概念，并以 0-10 为单位，以列表方式列出其相关度及相关度说明		
输出	核心概念	相关度(0-10)	相关度说明
	信息组织与检索	10	组织和检索信息资源是图书馆学的核心，直接影响资源可用性和用户体验。
	馆藏发展与管理	9	发展和管理丰富多样的馆藏以满足用户需求是图书馆的重要职责。
	读者服务	9	为读者提供优质服务，创造良好的用户体验，是图书馆吸引和留住用户的关键。
	信息服务	9	图书馆为用户提供的各种与信息相关的服务，包括参考咨询、用户教育、信息检索、文献传递等。
	信息资源	9	图书馆工作的对象，包括各种形式的文献、数据库、数字资源等，以及获取信息的途径。

示例 2：

提示	作为图书馆学专家，请将信息组织以递归方式列出其二级、二级类名
输出	信息组织 •方法 　•分类法 　　•杜威十进分类法 (DDC) 　　　•主类 　　　•次类 　　•美国国会图书馆分类法 (LCC) 　　　•主类 　　　•次类 •主题词表 　　•通用主题词表 　　•专业主题词表 •元数据

	•MARC •书目记录 •馆藏记录 •Dublin Core •简单 Dublin Core •复杂 Dublin Core

示例 3：

提示	作为图书馆学专家，请列出图书馆学最核心的 10 个概念元素，并以三元组形式输出，格式为：（<实体 1>，<关系>，<实体 2>）
输出	（图书馆学，研究对象，信息组织） (图书馆学，研究对象，信息检索) (图书馆学，研究对象，信息服务) (图书馆学，核心元素，信息资源) …

示例 4：

提示	作为图书馆学专家，请以递归方式列出图书馆学、信息组织及其二级、三级类名的属性三元组
输出	（图书馆学，研究对象，信息组织） •（信息组织，作用，管理信息资源） •（信息组织，目标，优化信息检索） •（信息组织，涉及，分类） •（分类，目标，组织信息资源） •（分类，使用，分类法） •（分类法，采用，十进分类法） •（分类法，采用，图书馆分类法） •（分类，采用，十进分类法） •（分类，采用，图书馆分类法） …

3. 实验及结果

实验围绕图书馆学领域知识的定义、采集、处理、抽取、融合及存储等环节展开，旨在探索领域知识图谱构建方法的普适性。

实验选取了四款最具代表性的商用语言模型（见表 1），这些模型在自然语言处理、对话系统、文本生成等多个领域表现出色,代表了当前大型语言模型的最高水平。

表1 构建领域知识图谱工具列表

Table 1 List of tools for constructing domain knowledge graphs

名称	功能
GPT-4o	多模态大模型，可以接受文本、音频和图像的任意组合作为输入内容，并生成文本、音频和图像的任意组合输出内容；可以进行数据分析、图像分析、互联网搜索、访问应用商店等操作。
Claude 3-opus	能够处理复杂分析、多步骤的长任务和高阶数学和编码任务，缩短上下文窗口，以优化速度和成本
Claude 3.5-Sonnet	能处理自主编码和视觉处理等复杂任务，并对于长文档处理上具有出色的优势，可确保检索增强生成（RAG）、搜索和检索以及比较多个长文档等任务的准确性。
Gemini-1.5-Pro	多模式模型，接受来自整个对话的文本、图片和视频输入并提供文本输出，但每条消息限制为一个视频。上下文窗口已缩短，以优化速度和成本。
JSON	用来存储简单的数据结构和对象的文件，可以在web应用程序中进行数据交换。
Gephi	一款开源免费跨平台基于JVM的复杂网络分析软件,，其主要用于各种网络和复杂系统，动态和分层图的交互可视化与探测开源工具。

（注：工具描述来源语言模型官方简介）

3.1 图书馆学领域本体构建

（1）图书馆学领域核心概念及等级体系划分

本文采用了不同的提示词设计，分别从GPT-4o、Claude-3-Opus、Claude-3.5-sonnet和Gemini-1.5-pro四种商用语言模型中抽取图书馆学领域核心概念。实验设置以10、15、20为单位递增的核心概念数量，并以0-10的相关度评分进行评估，结果显示，即使增加核心概念数量，相关度评分仍然集中在6分左右，并未呈现正态分布。因此，最终确定以10个最相关的核心概念为研究对象。

经过多次提示词文本设计和四种语言模型知识抽取，将获得的结果进行分析发现，“信息组织与检索”信息检索”作为图书馆学的研究对象并行存在，可将其分解为两个独立概念，根据“读者服务”与“用户体验”的内容描述，可整合为“信息服务”；根据“图书馆技术与自动化”的内容，可统一为信息系统。“信息素养教育”、“数字图书馆”、“知识管理”和“信息政策”因其相关度评分有所差异，则根据四次抽取的相关度均值进行确认，从而基本完成图书馆学领域核心概念的构建（见表2）。

表2 图书馆学领域核心概念类及描述

Table 2 Core conceptual categories and descriptions of Library Science

核心概念类名	描述
图书馆学	研究图书馆事业发展规律的科学，顶层概念类，本体中的所有类名均属于其子类
信息组织	对信息进行分析、标识、描述、排序和索引，以便于用户检索和利用的过程。
信息检索	根据用户表达信息需求，并从信息系统中获取所需信息的过程。
信息服务	提供各种形式的咨询、借阅、参考等服务，满足读者的信息需求。注重提升用户在使用图书馆资源和服务过程中的满意度和体验。
信息资源	科学地选择、采购、组织和维护图书馆的各类资源。根据用户需求和发展趋势，不断更新和优化馆藏结构。
信息素养教育	培养用户获取、评估、使用和创造信息的能力。通过各种形式的培训和指导，提高用户的信息意识和利用效率。
数字图书馆	利用数字技术构建的虚拟图书馆，提供电子资源的存储、组织和访问服务。扩展传统图书馆的功能，实现信息资源的广泛共享和远程访问。
知识管理	系统地收集、组织、分享和应用组织内的知识资源。促进知识的创新和传播，提高组织的学习能力和竞争力。

信息系统	应用各种信息技术和自动化系统，提高图书馆管理和服务的效率。
信息政策	制定和实施有关信息获取、使用和传播的政策和规范。关注信息使用中的道德问题，保护知识产权和用户隐私。
用户研究	通过各种方法调查和分析用户的信息需求、行为和满意度。为图书馆的服务策划和资源建设提供决策依据。
图书馆建筑	设计和建造适合图书馆功能的物理空间。考虑资源存储、用户服务、技术应用等需求，创造舒适、高效的学习和研究环境。

表 3 同样采取相关度评分的方式，对商用语言模型进行多次提示词设计，识别核心概念的二级、三级概念类别，为确保聚类集中，二级、三级概念的数量限制为 5 个。

表 3 图书馆学二级、三级概念类（部分）

table 3 Secondary and tertiary concepts of Library Science

上级核心概念类名	二级概念类	三级概念类
信息组织	知识发现	数据挖掘
		文本分析
		语义网
		机器学习
		主题分析
	用户导向	用户需求
		个性化服务
	编目	元数据
		书目记录
		著录规则
		著录格式
		自动化
	分类	图书馆分类法
	索引	文献内容
		标目
		索引词
		位置信息
		参考咨询
		自助检索

(2) 定义类的属性。

图 2 展示了利用提示词设计，从大语言模型中抽取图书馆学领域概念，并以三元组形式进行表示的部分结果。

(图书馆学 研究对象 信息组织)
(信息组织 目标 知识发现)
(知识发现 目标 揭示知识关联)
(知识发现 工具 数据挖掘)
(知识发现 工具 文本分析)
(知识发现 工具 社交网络分析)
(知识发现 技术 语义网)
(知识发现 技术 机器学习)
(知识发现 应用 主题分析)
(知识发现 应用 趋势预测)
(知识发现 应用 个性化推荐)
(信息组织 原则 用户导向)
(用户导向 强调 以用户为中心)
(用户体验 评估方法 可用性测试)
(用户体验 评估方法 满意度调查)
(用户参与 方式 志愿者服务)
(用户参与 方式 共建共享)
(用户导向 目标 满足用户需求)
(用户需求 获取方式 用户调查)
(用户需求 获取方式 数据分析)
(图书馆服务 原则 平等服务)
(图书馆服务 原则 个性化服务)
(图书馆空间 设计理念 以人为本)
(图书馆 发展趋势 用户服务智能化)
(图书馆 发展趋势 服务模式多元化)
(图书馆 发展趋势 馆员角色转变)
(信息组织 方法 编目)
(编目 工具 元数据)
(图书馆编目 目的 创建检索点)
(图书馆编目 对象 各种类型文献)
(编目规则 规定 著录规则)

图 2 图书馆学领域知识图谱本体（三元组）（部分）

Fig.2 Knowledge graph triple ontology of Library Science (part)

通过上述多次少样本提示词设计，从四种语言模型中抽取图书馆学领域的概念、二级和三级类别及其属性，形成该领域的知识集合。在整个抽取过程中，我们多次运用融合方法对本体及关系概念进行消歧，最终完成图书馆学领域知识图谱的本体构建。

3.2 知识图谱存储与构建

本研究的数据存储采用 JSON 文件格式。将知识融合后的本体及其属性三元组集，通过语言模型 Claude-3.5-Sonnet 转换为.GEXF 格式的图谱文件，并使用 Gephi 软件进行知识图谱的可视化。

提示词文本设计：

将<tag></tag>中的文字转换为一个.GEXF 格式的图谱文件，第一个元素和第三个元素作为节点，第二个元素作为边。<tag></tag>内嵌入需要转换的三元组集，如图 3。

```
<?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://www.gexf.net/1.2draft" version="1.2">
  <graph mode="static" defaultedgetype="directed">
    <nodes>
      <node id="图书馆学" label="图书馆学" />
      <node id="信息组织" label="信息组织" />
      <node id="知识发现" label="知识发现" />
      <node id="揭示知识关联" label="揭示知识关联" />
      <node id="数据挖掘" label="数据挖掘" />
      <node id="文本分析" label="文本分析" />
      ......
    </nodes>
    <edges>
      <edge source="图书馆学" target="信息组织" label="研究对象" />
      <edge source="信息组织" target="知识发现" label="目标" />
      <edge source="知识发现" target="揭示知识关联" label="目标" />
      <edge source="知识发现" target="数据挖掘" label="工具" />
      <edge source="知识发现" target="文本分析" label="技术" />
      ........
    </edges>
  </graph>
</gexf>
```

图 3　图书馆学领域本体图谱文件

Fig. 3 Map file of Library Science domain ontology

图谱文件中，知识本体作为节点，存储为“实体 id，实体名称，label”，关系或者属性作为边，存储为“源实体 source，目标实体 target，类型为有向”，通过 Gephi 可视化分析软件，形成图书馆学领域知识图谱，其中包括 173 个节点，145 个边。图 4 为图书馆学领域知识图谱总览。

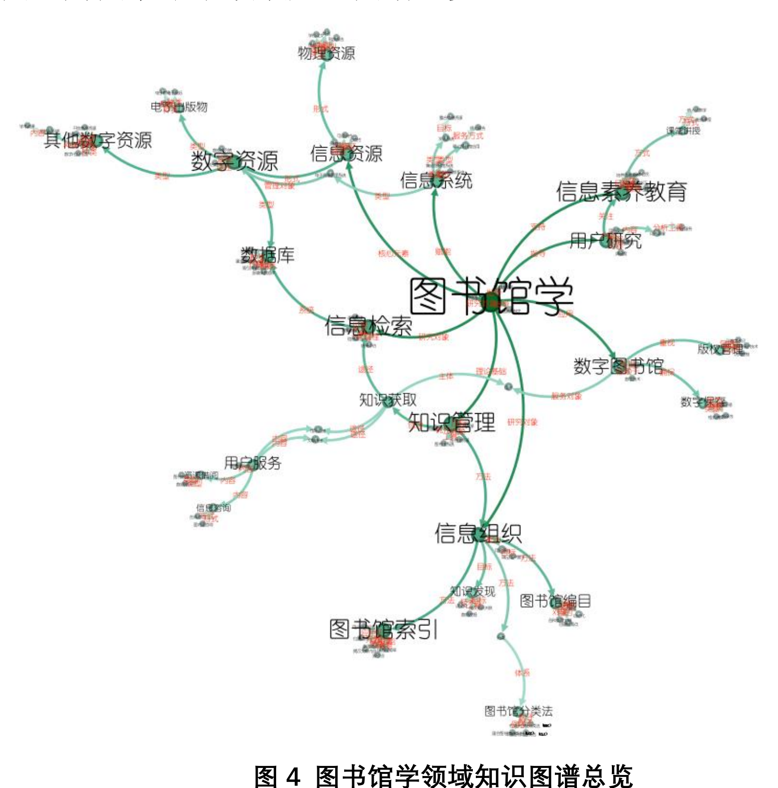


图 4　图书馆学领域知识图谱总览

Fig. 4 Knowledge graph overview of Library Science

4. 总结

相较于传统的知识图谱构建方法，基于大预言模型的少样本学习在构建知识图谱方面具有显著优势。它不仅能大大降低了人工标注成本，提高构建效率，还能补全传统构建方法的不足，增强构建知识图谱的可扩展性和泛化能力。此外，大语言模型还能实现高度自动化的构建流程，减少人工干预，提高构建的一致性。凭借强大的文本处理能力和深度理解能力，大语言模型可以挖掘实体之间潜在的隐含关系，构建更全面、更丰富的知识图谱。

然而，基于大语言模型的少样本学习方法也存在一些问题。例如，模型的学习效果很大程度上依赖于提示词的设计，设计不当可能导致模型无法理解任务或生成有偏差的结果。模型对领域知识的抽取，尤其是在特定领域，可能存在偏差，需要领域知识专业词典的补充等问题。

参考文献：

[1] OpenAI. (2023). *ChatGPT-4 Technical Report*. Retrieved from https://openai.com/research/gpt-4

[2] 王娟,曹树金,王志红,彭碧涛.面向探索式搜索的领域知识图谱构建及实验探索[J].图书情报工作,2024,68(03):105-116.

[3] Ehrlinger, M., Wöß, W. Towards a Definition of Knowledge Graphs[J]. Datenbank-Spektrum, 2016,16(1): 9-19.

[4] 汪莉,姜楠,程斌,董昌武.知识图谱在中医药领域应用的研究概况[J].西安文理学院学报(自然科学版),2024,27(03):70-75.

[5] Paulheim, H. Knowledge graph refinement: A survey of approaches and evaluation methods [J]. Semantic Web, 2016, 8(3): 341-370.

[6] Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. A survey on knowledge graphs: Representation, construction, and applications [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2), 494-514.

[7] Wang, Y., Li, M., & Wang, J. BioNER: A CRF-based Biomedical Named Entity Recognition Systemv [J]. Journal of Biomedical Informatics,2009,42(4), 676-685.

[8] Lample, G., Ballesteros, M., Kawakami, K., Subramanian, S., & Dyer, C. Neural Architectures for Named Entity Recognition[C]//*Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260-270.

[9] Angles, R., & Gutierrez, C. Survey of Graph Database Models[J]. ACM Computing Surveys, 2008,40(1), 1-39.

[10] 黄勃,吴申奥,王文广,等.图模互补：知识图谱与大模型融合综述[J/OL].武汉大学学报(理学版),2024,41(04):397-412.https://doi.org/10.14188/j.1671-8836.2024.0040.

[11] 陈昱成,黎洋,刘江峰,等.AIGC 视角下非物质文化遗产知识图谱的构建研究[J].科技情报研究,2024,6(02):115-128. https://doi.org/10.19809/j.cnki.kjqbyj.2024.02.010.

[12] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing [J]. ACM Computing Surveys, 2023, 55(9), 1-35.

[13] Wang, Q., Wang, B., Guo, L., Zhang, Z., Wang, X., & Han, J. Knowledge-enhanced pre-trained language models for natural language processing: A survey [J]. arXiv preprint

arXiv:2020,2009.08854.
[14] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. Language models as knowledge bases? [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,2019,2463–2473.
[15] Ding, H., Zhou, P., Huang, M., & Zhu, X. Exploiting Cloze-style Tasks for Emotion Cause Extraction[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020,5409–5419.
[16] Yao, L., Mao, C., & Luo, Y. KG-BERT: BERT for Knowledge Graph Completion[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020,34(05), 9487–9494.
[17] Shin, T., Raffel, C., Shah, U., & Roberts, A. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020,4222–4235.
[18] Lv, X., Wang, Y., Zhu, H., Sun, M., & Huang, F. (2020). Meta Relational Learning for Few-Shot Link Prediction in Knowledge Graphs [J]. arXiv preprint arXiv:2020,2006.08198.